

Title:

Item Response Theory

Authors:

Li Cai

Kilchan Choi

Mark Hansen

Lauren Harrell

Journal publication date:

2016

IES grant information:

Grant number R305D140046

Funded by National Center for Education Research (NCER)

## Item Response Theory

Li Cai

with additional contributions by

Kilchan Choi

Mark Hansen

Lauren Harrell

University of California, Los Angeles

Oct 7, 2015

Chapter contribution to *Annual Review of Statistics and Its Application*

The author's research is supported in part by the Institute of Education Sciences (R305D140046). The views expressed here belong to the author and do not reflect the views or policies of the funding agency. The author wishes to thank Dr. Wenjing Huang for helpful comments and edits.

Address all correspondence to: Li Cai, CRESST, UCLA, Los Angeles, CA, USA 90095-1521.  
Email: [lcail@ucla.edu](mailto:lcail@ucla.edu). Phone: 310.206.0583. Fax: 310.206.5830.

## 1 Introduction

To the statistician, Item Response Theory (IRT) typically represents a part of discrete multivariate analysis focused on modeling correlated multivariate response variables in a structured (often hierarchical) manner in terms of underlying latent variables of either continuous (latent factors) or discrete (latent classes) nature. IRT is frequently used in educational or psychological testing or patient-reported health outcomes, where data from questionnaires or standardized measurement instruments are modeled. In this review we shall keep these areas foremost in mind, while trying to at the same time to place the material in a broader statistical context.

In a more pragmatic sense, IRT encompasses a set of models along with the requisite computational and inferential tools that help researchers address such technical issues as item analysis, score reliability, scale alignment, etc., that are related to the inherent fairness, quality, and validity questions (e.g., as defined by the joint standards on testing; AERA, APA, NCME, 2014) associated with the development, administration, maintenance, interpretation, and use of tests. Particularly salient features of IRT models include the use of latent variables to represent the constructs being assessed such as mastery, achievement, attitude, or severity of a disorder; the direct use of item-level (mainly categorical) data collected from the interactions of individual examinees, respondents, or patients with instances of assessments; as well as the reliance on conditional independence as the basic structure and recipe for model-building.

As Bock (1997) and Thissen and Steinberg (2009) noted, IRT has enjoyed continued development in the field of psychometrics, with many hundreds of papers and reports written over the last eight or nine decades, appearing frequently in such outlets as *Psychometrika*, *Journal of Educational and Behavioral Statistics*, and *Journal of Educational Measurement*, among others. Book-length volumes on IRT, ranging from introductory (e.g., Hambleton, Swaminathan & Rogers, 1991) to more technical (e.g., Baker & Kim, 2004), have been published. In many cases, the books have also witnessed revisions over the years to reflect growing developments in the research literature.

It is fair to say that IRT is now one of the central methodological pillars supporting many large and high-profile assessment programs around the globe, e.g., OECD's Program for International Student Assessment, the National Assessment of Educational Progress (a.k.a. the Nation's Report Card in the US), National Institutes of Health's Patient Reported Outcomes Measurement Information System (PROMIS®) initiative, China's new National Assessment of Basic Education Quality, and the list goes on. It has also become increasingly applied in academic research settings to enhance the precision and rigor of measurement in social and behavioral sciences (e.g., Reise & Waller, 2009).

Most of the applications of IRT methods to date are restricted to the case of unidimensional latent trait models for single level data. That is, most of the time, researchers assume the presence of a single continuous underlying latent variable that explains the covariation of the observed item responses. At the same time, the item response vectors are assumed to be independent across individuals. Hierarchically nested structures, such as students nested within classrooms, or repeated measures nested within persons are often ignored. Unfortunately, many constructs in the social and behavioral sciences are multifaceted (for example, a chronic patient's quality of life), calling for multidimensional item response models. Multilevel item response models, on the other hand, provide mechanisms to represent the extra dependence introduced by nested data structures. Furthermore, trait-like continuum representing individual differences is not the only kind of latent variables that are theoretically plausible and interpretable. For instance, models containing latent classes or mixtures of classification and continuous latent variables are needed in applications of IRT.

Because of IRT's disciplinary focus on the social sciences and its association with educational and psychological measurement in particular, the vocabulary and mathematical notation in primary sources can sometimes seem esoteric to the general applied statistical audience. This review will attempt to avoid that by using more customary notation (e.g.,  $\theta$  refers to parameters) and by beginning the discussion with a few motivating examples so that key concepts can be introduced.

## 2 Some Motivating Examples

### 2.1 A Beginning Example: Assessment Items in Education or Psychology

The item characteristic curve is one of the fundamental building blocks in the machinery of IRT. The item characteristic curve describes the conditional probabilities of observed item responses (e.g., correct vs. incorrect response) given unobserved characteristics. In educational and psychological measurement, one of the most frequently encountered item characteristic curve is the 3-parameter logistic (3PL) function used in this connection by Birnbaum (1968):

$$T_i(1|\eta) = \gamma_i + \frac{1 - \gamma_i}{1 + \exp[-(\alpha_i + \beta_i\eta)]} \quad (1)$$

where the  $T_i$  notation pays homage to the earlier work on IRT by Paul Lazarsfeld (1950), who called the item characteristic curves tracelines because they “trace” the observable probability of a “correct” or “endorsement” response given unobserved variable  $\eta$ . The 3 parameters  $\alpha_i$  (intercept),  $\beta_i$  (slope or discrimination), and  $\gamma_i$  (pseudo-guessing parameter) are specific to item  $i$ , and as such they are referred to as item parameters, in contrast to those distributional parameters that are associated with the specification of the prior density of the latent variable  $\eta$ , e.g., the mean and variance of  $\eta$ . If a random sampling interpretation is taken, the distribution of  $\eta$  can also be called the population density. The intercept is inversely associated with the degree of difficulty of the item. Because the rate of change in correct response probability moves faster for an item with a larger slope parameter, the slope is directly related to how well the item can differentiate among people having different degrees of proficiency. The pseudo-guessing parameter reflects the non-zero probability of observing correct answers even for students with low proficiency (e.g., by randomly guessing). In psychological assessments, when IRT is used to model item about symptoms of a psychological disorder, it is also possible that some items will be endorsed even when the individual’s overall symptom severity is low, potentially due to other unmodeled external causes (see e.g., Reise and Waller, 2009). Computationally and statistically efficient estimation of item and other distributional parameters as well as their standard errors is central to IRT modeling.

The 3PL function shown in Equation (1) has straightforward interpretations. For the purpose of illustration, consider the following example. A total of 5,750 students in an elementary grade took a set of state mathematics assessment items being field tested. There were 79 multiple-choice items, scored dichotomously as correct (1) vs. incorrect (0). The panel on the left in Figure 1 shows the estimated 3PL item characteristic curve for item #75. The underlying latent variable could be understood as math proficiency. By convention in IRT, one usually fixes the location and scale of the latent variable by imposing a standard normal prior on  $\eta$ . With the method of maximum marginal likelihood (Bock & Aitkin, 1981), the intercept is estimated to be  $-.18$  ( $SE = .13$ ), with a slope of  $1.46$  ( $SE = .11$ ), and the pseudo-guessing probability is  $.28$  ( $SE = .03$ ). As one can see, the curve takes on a sigmoidal shape, with a noticeable non-zero lower asymptote. Item #75 is not overly difficult as it has a middling intercept value. From experience, item #75 is also a reasonably discriminating item. The presence of substantially non-zero correct response probability toward the lower end of the summed score scale indicates the effect of guessing.

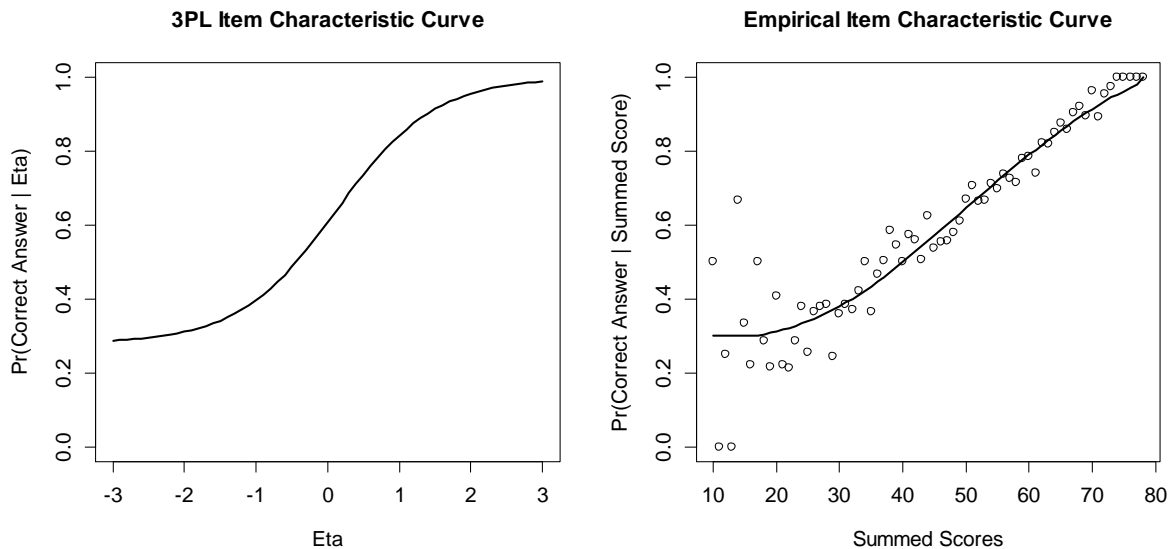


Figure 1. Example Item Characteristic Curve with Its Empirical Counterpart

Simpler models for dichotomous responses do exist. For instance, the lower asymptote may be constrained to 0, which results in the so-called 2-parameter logistic (2PL) model. The item slopes may also be constrained to equality across items, resulting in the 1-

parameter logistic (1PL) model, which is the identical twin of the Rasch model (Rasch, 1961). One could also use altogether different link functions, e.g., the probit, to derive the more traditional normal ogive IRT models, instead of their logistic counterparts. The motivation and derivation of these various models may be pursued either along the lines of the Rasch measurement tradition or as direct descendants and extensions of Thurstone's (1925) models (see e.g., Thissen & Wainer, 2001; Thissen & Steinberg, 2009).

Yet perhaps a more direct rationale for the use of these sigmoidal item response functions can be gleaned from an entirely empirical observation that should appeal to the applied statistician. The panel on the right in Figure 1 shows an empirical item characteristic curve, formed by tabulating the raw item response data. Setting item #75 aside, there are 78 other items. Students' observed total (summed) item scores may serve as a close proxy (assuming that the latent variable is unidimensional) for the unobserved proficiency variable  $\eta$ . The probabilities of responding correctly within each of the distinct summed score groups (ranging from 10 to 78 in this case), when plotted against the summed scores, follow a familiar S-shaped curve. For very low summed scores, the frequencies in those groups are also small, resulting in more variability at the lower end. What is superimposed on the data points is the 3PL model-implied correct response probabilities for each of the summed score groups, based on the estimated item parameter values shown earlier. These probabilities are computed using a recursive algorithm first described by Lord and Wingerksy (1984). Plots and summaries such as these not only helpful as empirical motivations for IRT modeling, but also form the basis of a number of model fit diagnostic indices (see Cai, 2015). Extending this example, more complex models and applications will be reviewed, moving from dichotomous variables to polytomous variables, unidimensional to multidimensional, and from single level to multilevel.

## 2.2 Large-Scale Educational Surveys

Large-scale educational surveys play important roles in research and policy discussions. In the United States, people carefully watch over long-term trajectories and changes in student achievement as measured by the National Assessment of Educational Progress (NAEP), a Congressionally-mandated program. Across national boundaries, high-

profile rankings based on Program for International Student Assessment (PISA), commissioned by the Organization for Economic Cooperation and Development (OECD), have sparked heated debates in the last decade about national educational competitiveness.

In order to support NAEP, research conducted at the Educational Testing Service (ETS) in the 1990's, led by pioneers such as Robert Mislevy, among others, resulted in a general approach that is based on IRT and the plausible value (PV) methodology. The PV methodology is derived from Rubin's (Rubin, 1987; Little & Rubin, 2002) multiple imputation theory for analysis involving missing data. In NAEP, student achievement outcomes are considered latent variables and hence they are unobserved/missing data. Student responses to assessment items and individual background information such as demographic variables and upper-level characteristics (e.g., at the school level) serve the role of observed data. To minimize respondent burden and cost, and more importantly, to provide efficient estimates of means, variance components, coefficients, and contrasts at levels of aggregation (e.g., districts, states, and countries in the case of PISA) where principal policy interests lie, the design of large-scale educational assessments as exemplified by NAEP also makes use of sampling of items so that each student only receives a small number of items, resulting in high fractions of missing data at the individual level. The problem is that estimates of the population parameters are biased when individual level fractions of missing information are high (Mislevy, 1991).

Rather than a single point estimate of student-level proficiency, multiple PVs are "imputed" from a posterior distribution that includes the student's item responses and background characteristics (Mislevy, 1991; Mislevy, Johnson, & Muraki, 1992). Central to the PV methods is a latent regression model that has components from both IRT and the general linear model.

Let  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$  be the matrix of background covariates, where  $\mathbf{x}_j$  is the covariate vector for individual  $j$ , and  $N$  is the total sample size. The background covariates may include information about demographics, schooling characteristics, motivation, resources, peer contexts, etc. Let  $\mathbf{y}_j = (\mathbf{y}_{j1}, \dots, \mathbf{y}_{jD})$  be the individual assessment outcomes on  $D$  domains of interest (which are principally defined by policy interests, e.g., reading, math,



and science for PISA), comprised of item responses  $\mathbf{y}_{jd}$  with typical element  $y_{ijd}$  denoting item  $i$  within domain  $d$ , and  $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$  collects together all observed outcomes. Of interest is estimating the unobserved student achievement in  $D$  domains, and the vector of these latent variables for student  $j$  is represented by  $\boldsymbol{\eta}_j = (\eta_{j1}, \dots, \eta_{jD})$ .

In modeling student responses, multiple categorical items are frequently encountered in addition to the dichotomously scored multiple-choice items. The multiple categorical responses may result from ordinal ratings of constructed responses such as essays or short answers. The generalized partial credit (GPC) model (Muraki, 1992) is a widely used model for such item responses. Suppose item  $i$  has  $C_i$  categories, with item codes  $c \in \{0, \dots, C_i - 1\}$ . A version of the GPC model may be written as

$$T_i(c|\boldsymbol{\eta}) = \frac{\exp(c\beta_i\boldsymbol{\eta} + \alpha_{ic})}{\sum_{m=0}^{C_i-1} \exp(m\beta_i\boldsymbol{\eta} + \alpha_{im})}, \quad (2)$$

where the overall item slope  $\beta_i$  serve the same role as the discrimination parameter of a 3PL model. The category-specific intercept  $\alpha_{ic}$  parameters are not identified without further constraints, typically with the first intercept fixed to 0 ( $\alpha_{i0} = 0$ ). More broadly, the GPC model is one of a large number of models in the nominal categories model family (for details, see Thissen, Cai & Bock, 2010).

In the NAEP conditioned latent regression model that produces the plausible values, the latent proficiency vector  $\boldsymbol{\eta}_j$ , is assumed to be normally distributed with conditional mean vector  $\boldsymbol{\Gamma}\mathbf{x}_j$  and covariance matrix  $\boldsymbol{\Sigma}$ , where  $\boldsymbol{\Gamma}$  represents a matrix of unknown regression coefficients (Thomas & Gan, 1997). The background covariates contribute to the explanation of the individual-specific outcomes by giving each student potentially different conditional means. Let  $f(\boldsymbol{\eta}_j; \boldsymbol{\Gamma}\mathbf{x}_j, \boldsymbol{\Sigma})$  represent a multivariate normal density function with mean  $\boldsymbol{\Gamma}\mathbf{x}_j$  and covariance matrix  $\boldsymbol{\Sigma}$ . Suppose that the items in each of the proficiency domains are independent once the latent proficiency  $\eta_{jd}$  is accounted for – this is known as the independent cluster factor pattern in the factor analysis literature. The posterior distribution of the latent proficiencies  $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_N)$  is then proportional to

$$p(\boldsymbol{\eta}|\mathbf{X}, \mathbf{Y}) \propto \prod_{j=1}^N f(\boldsymbol{\eta}_j; \boldsymbol{\Gamma}\mathbf{x}_j, \boldsymbol{\Sigma}) \prod_{d=1}^D f_d(\mathbf{y}_{jd}|\eta_{jd}), \quad (3)$$

from which PVs of  $\boldsymbol{\eta}$  can be drawn.

The current NAEP “machinery” is not without its own limitations. The current framework defines the proficiency domains *a priori*, and fails to account for any conditional dependence between items on different proficiency exams such as the effect of testlets (Cai, 2010a). The current formulation also does not allow for investigators to explore alternative parameterizations of the latent proficiency space, e.g., combining items from different domains. Fully multidimensional IRT models with simultaneous estimation is one avenue to resolve these difficulties.

### 2.3 Quantifying Effectiveness in Multisite Randomized Trials

Over the last decade or more, American education research policy took a dramatic shift toward the use of randomized experiments to evaluate the effectiveness of intervention programs. In a review of US Department of Education funded trials, the majority were multisite cluster randomized experiments (Spybrook & Raudenbush, 2009). In multisite cluster randomized trials, clusters (e.g., intact classrooms) within a site (e.g., schools) are randomly assigned to different conditions. Among its many benefits, the multisite trial essentially entails the use of “mini-experiments” at each site, which can be helpful in examining the generalizability of experimental results (Raudenbush & Liu, 2000).

Multisite cluster randomized designs in the social and behavioral sciences typically include a test-retest administration of (almost always) identical assessments of the outcome construct to the same group of individuals at two time points, before and after the intervention. For example, the outcome could be a measure of math proficiency that the intervention is supposed to influence at a particular grade of interest.

In practice, the most commonly used method for scoring outcome measures is with the summed score. That is, the total score from an assessment instrument is computed and used as observed outcome or predictor values. Researchers have also increasingly utilized IRT based approaches (Curran et al., 2008) and using estimated individual scaled scores such as

the Expected *A Posteriori* (EAP) estimates (Thissen & Wainer, 2001). The IRT scaled scores are then treated as continuous outcome or predictor variables in subsequent data analyses.

The EAP method is not difficult to understand. Given estimated item parameters, the shape of the item characteristic curves are assumed pre-determined. With the strong assumption of conditional independence of the item responses, one may write the joint probability of observing the individual vector of item responses  $\mathbf{y}_j = (y_{1j}, \dots, y_{nj})$  as

$$f(\mathbf{y}_j|\eta) = \prod_{i=1}^n f(y_{ij}|\eta), \quad (4)$$

where the conditional probability mass function of an individual item response is that of a multinomial with trial size 1 in  $C_i$  categories,

$$f(y_{ij}|\eta) = \prod_{c=0}^{C_i-1} T_i(c|\eta)^{\chi_c(y_{ij})}, \quad (5)$$

with the indicator function

$$\chi_c(y_{ij}) = \begin{cases} 1, & \text{if } c = y_{ij} \\ 0, & \text{otherwise} \end{cases}. \quad (6)$$

Here, the assumption of conditional item independence is explicit and the posterior of the latent variable  $\eta$  is

$$f(\eta|\mathbf{y}_j) = \frac{\prod_{i=1}^n f(y_{ij}|\eta) \phi(\eta)}{\int \prod_{i=1}^n f(y_{ij}|\eta) \phi(\eta) d\eta}, \quad (7)$$

where  $\phi(\eta)$  is the prior (population) distribution of the latent variable. While  $f(\eta|\mathbf{y}_j)$  must typically be approximated numerically (with quadrature), the computation of the posterior mean and variance is straightforward. Importantly, the EAP estimator shrinks individual scaled scores toward a common grand mean, as specified in the population distribution.

For a number of reasons, the existing IRT scaled score computation methods are problematic for randomized multisite intervention studies, despite being straightforward. First, the outcome constructs at each occasion are correlated due to the longitudinal design. The dependence of latent constructs pre- and post-treatment makes the choice of appropriate sample for item analysis ambiguous. Second, there is likely substantial item-level residual dependence due to repeated (pre-post) administration of the same items to the

same individuals. The item-level residual dependence, if unattended to, might lead to overly optimistic estimate of statistical uncertainty. Third, the individuals are nested in sites, so there is likely extra correlation that should be modeled. Failure to account for such nesting may lead to the lack of congruence between statistical models used for measurement (single-level IRT model) and impact estimation (typically multilevel hierarchical model). Last but not the least, it is implausible to assume full exchangeability of individuals across treatment and control conditions. That is, the random variables associated with post-treatment individuals in the control condition cannot be exchanged without consequences with those in the treatment condition (see e.g., Lindley & Smith, 1972). What is needed in scoring is to assume conditional exchangeability, conditioning on the treatment assignment indicator, which is unfortunately something standard EAP calculations do not account for.

The last issue of exchangeability is particularly important, because failure to include important conditioning information (e.g., treatment assignment) in the measurement model can and will lead to inconsistencies in the subsequent inference about treatment effect. A more direct approach proposed by Cai, Choi, and Kuhfeld (in press), using a multilevel extension of Cai's (2010a) two-tier item factor analysis model, will be illustrated later in this chapter as a tool that systematically addresses these measurement issues in randomized evaluation studies of the efficacy of interventions.

## 2.4 Diagnostic Measurement

Diagnostic classification models (DCMs; see, e.g., Rupp, Templin, & Henson, 2010) have received increasing attention within the field of educational and psychological measurement. In contrast to more traditional IRT models discussed thus far where continuous latent variables are specified, DCMs presupposes the presence of discrete latent variables/attributes (e.g., mastery of skills or knowledge, presence of symptoms, possession of misconceptions, etc.) that explain the observed covariation of the item responses. For simplicity, consider a  $Q \times 1$  vector of dichotomous (0-1) underlying latent variables,  $\xi$ . The relationship between items and latent attributes is captured by the so-called **Q**-matrix (Tatsuoka, 1983), which is a  $n \times Q$  matrix of fixed zeros and ones. The  $(i, q)$ th entry in the **Q**-matrix takes on a value of 1 if item  $i$  measures attribute  $q$ .

Let  $T_i(c|\xi)$  be the response function for item  $i$  in category  $c \in \{0, \dots, C_i - 1\}$ . A useful model to consider for ordered polytomous item responses is cumulative logit model based on Samejima's (1968) original graded response model. Let

$$T_i^+(c|\xi) = \frac{1}{1 + \exp(-z_{ic})}, \quad (8)$$

be the conditional cumulative response probability in categories  $c$  and above. Then the category response probability is the difference

$$T_i(c|\xi) = T_i^+(c|\xi) - T_i^+(c+1|\xi), \quad (9)$$

where  $T_i^+(0|\xi) = 1$  and  $T_i^+(C_i|\xi) = 0$  are the fixed boundary cases. In the case of DCMs, the linear predictor of the graded response model  $z_{ic}$  is

$$z_{ic} = \alpha_{ic} + \beta_i' h_i(\mathbf{Q}, \xi). \quad (10)$$

It is easy to see that the model contains  $C_i - 1$  intercepts ( $\alpha_{i1}, \dots, \alpha_{i,C_i-1}$ ). The  $h_i(\mathbf{Q}, \xi)$  term in Equation (10) is a potentially vector-valued function that defines how the measured attributes combine to create the linear predictor portion of the item response model in Equation (8). As noted by Henson, Templin and Willse (2009), constraints on the item parameters and choices of  $h_i(\mathbf{Q}, \xi)$  yield several of the more commonly utilized diagnostic models (see also Rupp et al., 2010; Choi, Rupp, & Pan, 2013).

For instance, suppose that according to the  $\mathbf{Q}$ -matrix, a mathematics test item  $i$  measures two particular algebra-related skills (attributes  $\xi_1$  and  $\xi_2$ ), and that successful solution of the item (so that the item score  $c = 1$ ) requires both attributes. The linear predictor may take the following deterministic-input noisy "and" gate (DINA; Junker & Sijtsma, 2001) form, with an intercept and a single free parameter for the interaction term:

$$z_{i1} = \alpha_{i1} + \beta_i \xi_1 \xi_2. \quad (11)$$

In this case,  $h_i(\mathbf{Q}, \xi) = \xi_1 \xi_2$  contains second-order interaction, and  $\beta_i = \beta_i$ . Alternatively, perhaps the item only requires the mastery of *either* attribute 1 or attribute 2. Then the linear predictor of the IRT model may take the following deterministic-input noisy "or" gate (DINO; Templin & Henson, 2006) form, with the slope set equal between the main effects and the interaction in magnitude, albeit the interaction has a negative sign:

$$z_{i1} = \alpha_{i1} + \beta_i \xi_1 + \beta_i \xi_2 - \beta_i \xi_1 \xi_2. \quad (12)$$

In this case  $\beta_i = (\beta_i, \beta_i, -\beta_i)'$  contains linear restrictions on the item parameters. Yet another possibility is that each attribute contributes to some increase in the log-odds of correct solution, and that the magnitude of the increase due to one attribute does not necessarily require the mastery of the other. In that case, the linear predictor might contain only the main effect terms (Hartz, 2002; von Davier, 2005), similar to factor analysis:

$$z_{i1} = \alpha_i + \beta_i \xi_1 + \beta_i \xi_2. \quad (13)$$

DCMs can also be further integrated with item factor analysis models to provide even more flexible frameworks for data analysis, which will be discussed next.

### 3 A More General Framework

Given the foregoing discussion, a general modeling framework will be presented as a systematic synthesis of recent research on multidimensional, multilevel, and diagnostic classification IRT models. Consider a response in observed category  $c$  to item  $i$ . Suppose the level-one unit (respondent)  $j$  is nested within independent sampling unit (level-two unit)  $k$ , e.g., a student nested within a classroom. The linear predictor can be written as

$$z_{ijkc} = \alpha_{ic} + \beta'_{i1} \begin{pmatrix} h_{i1}(\mathbf{Q}_1, \xi_{jk}) \\ \boldsymbol{\eta}_{jk} \end{pmatrix} + \beta'_{i2} \begin{pmatrix} h_{i2}(\mathbf{Q}_2, \xi_k) \\ \boldsymbol{\eta}_k \end{pmatrix}, \quad (14)$$

where  $\beta_{i1}$  and  $\beta_{i2}$  are the item slope parameters, with  $\mathbf{Q}_1$  and  $\mathbf{Q}_2$  the  $\mathbf{Q}$ -matrices at levels one and two, respectively. The subscripts on  $\xi$  and  $\boldsymbol{\eta}$  indicate whether the latent variables vary over level-one units ( $jk$ ) or over level-two units ( $k$  only). The dimensions of the slope vectors are dependent on both the number of continuous latent factors in  $\boldsymbol{\eta}$ , as well as the number of terms generated by the  $h$  functions at each level.

Substituting the linear predictor in Equation (14) into any one of the item response functions mentioned earlier (e.g., the 3PL), the conditional probability of observing response  $y_{ijk}$  remains that of a multinomial with trial size 1 in  $C_i$  categories (as in Equation 5):

$$f(y_{ijk} | \xi_{jk}, \xi_k, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_k) = \prod_{c=0}^{C_i-1} T_i(c | \xi_{jk}, \xi_k, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_k)^{\chi_c(y_{ijk})}, \quad (15)$$

where the indicator function  $\chi_c(y_{ijk})$  is similarly defined as in Equation (6). In order to build the model further, one needs to invoke the conditional independence assumption so

that the conditional response pattern probability for response vector  $\mathbf{y}_{jk} = (y_{1jk}, \dots, y_{njk})$  can be expressed as a product of individual item response probabilities.

$$f(\mathbf{y}_{jk} | \boldsymbol{\xi}_{jk}, \boldsymbol{\xi}_k, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_k) = \prod_{i=1}^n f(y_{ijk} | \xi_{ijk}, \xi_k, \eta_{ijk}, \eta_k). \quad (16)$$

Conditional independence is a strong assumption that deserves rigorous checking. In Section 5, model fit indices pertinent to examining conditional independence will be reviewed. Regardless, in contrast to unidimensional IRT models reviewed earlier, the presence of both types of latent variables (continuous and discrete) at both levels help better explain the covariation of item responses, making conditional independence more plausible in substantive measurement situations.

To complete the specification of the model, the relation among discrete latent attributes is further modeled in a higher-order factor model (see e.g., de la Torre & Douglas, 2004). This amounts to treating the  $\xi$ 's as though they were unobserved "items" and regressing them (first-order latent variables) on other continuous (second-order) latent factors. For example, for dichotomous attributes, a multidimensional extension of the 2-parameter logistic model (Reckase, 2009) may be used to relate the latent attributes to the higher-order latent factors:

$$T_q(1 | \boldsymbol{\zeta}_{jk}) = \frac{1}{1 + \exp[-(\tau_{q1} + \boldsymbol{\lambda}'_{q1} \boldsymbol{\zeta}_{jk})]} \quad (17)$$

where  $\tau_{q1}$  represents the latent attribute intercept and  $\boldsymbol{\lambda}_{q1}$  is a vector of attribute slopes that describe how associated each of the higher-order factors are related to the first-order attributes. Of course, Equation (17) is intended for latent variables at level one. For level-two attributes, we can define a similar expression:

$$T_q(1 | \boldsymbol{\zeta}_k) = \frac{1}{1 + \exp[-(\tau_{q2} + \boldsymbol{\lambda}'_{q2} \boldsymbol{\zeta}_k)]} \quad (18)$$

regressing the level-two attributes on higher-order latent factors. Again, if we assume conditional independence of the latent attributes given  $\boldsymbol{\zeta}$ , we may write

$$f(\boldsymbol{\xi}_{jk} | \boldsymbol{\zeta}_{jk}) = \prod_{q=1}^{Q_1} [T(1 | \boldsymbol{\zeta}_{jk})]^{\xi_{jkq}} [1 - T(1 | \boldsymbol{\zeta}_{jk})]^{1 - \xi_{jkq}}, \quad (19)$$

$$f(\xi_k | \zeta_k) = \prod_{q=1}^{Q_2} [T(1 | \zeta_k)]^{\xi_{kq}} [1 - T(1 | \zeta_k)]^{1-\xi_{kq}},$$

in order to fully represent the conditional attribute profile probabilities.

The key benefit to this higher-order factor model specification is that of simplicity. For  $q$  dichotomous attributes at either level, the unconstrained number of attribute profiles is equal to  $2^q$ , with a huge number of parameters ( $2^q - 1$ ) even if the number of attributes is only reasonably large (e.g., a dozen). With the higher-order model and a single higher-order latent factor, the number of parameters is substantially reduced to  $2q$ .

The remaining task involves the conditioning of the  $\eta$  and  $\zeta$  variables by regressing them on background covariates at levels one and two. Consider the latent regression of the  $\eta$  variables:

$$\begin{pmatrix} \eta_{jk} \\ \eta_k \end{pmatrix} = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \mathbf{0} & \Gamma_{22} \end{pmatrix} \begin{pmatrix} \mathbf{x}_{jk} \\ \mathbf{x}_k \end{pmatrix} + \begin{pmatrix} \epsilon_{jk} \\ \epsilon_k \end{pmatrix}, \quad (20)$$

where the lower left block of the regression coefficient matrix is constrained so that level-two latent variables are not directly regressed on level-one covariates. A similar equation can be specified for the higher-order factors  $\zeta$ :

$$\begin{pmatrix} \zeta_{jk} \\ \zeta_k \end{pmatrix} = \begin{pmatrix} \Pi_{11} & \Pi_{12} \\ \mathbf{0} & \Pi_{22} \end{pmatrix} \begin{pmatrix} \mathbf{x}_{jk} \\ \mathbf{x}_k \end{pmatrix} + \begin{pmatrix} \delta_{jk} \\ \delta_k \end{pmatrix}. \quad (21)$$

The equation disturbance terms have zero means and error covariance matrices that may contain estimable parameters subject to identification.

Putting various parts of the model together, and following standard modeling practices, the latent variables can be integrated out of the joint likelihood sequentially. Let  $\theta$  denote a  $d \times 1$  vector that collects together all free parameters in the model. These include parameters from all the items (the  $\alpha$ 's,  $\beta$ 's, and  $\gamma$ 's), parameters for the latent variables, the higher-order model (the  $\lambda$ 's), and the regression coefficients and error covariance matrices. In the first step, the level-one discrete latent attributes are integrated out:

$$f_{\theta}(\mathbf{y}_{jk} | \zeta_{jk}, \xi_k, \eta_{jk}, \eta_k) = \int f_{\theta}(\mathbf{y}_{jk} | \xi_{jk}, \xi_k, \eta_{jk}, \eta_k) f_{\theta}(\xi_{jk} | \zeta_{jk}) d\xi_{jk}. \quad (22)$$

Second, the level-one continuous latent variables are integrated out, bringing in the conditioning covariates as well:



$$\begin{aligned}
& f_{\theta}(\mathbf{y}_{jk} | \boldsymbol{\xi}_k, \boldsymbol{\eta}_k, \mathbf{x}_{jk}, \mathbf{x}_k) \\
&= \iint f_{\theta}(\mathbf{y}_{jk} | \boldsymbol{\zeta}_{jk}, \boldsymbol{\xi}_k, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_k) f_{\theta}(\boldsymbol{\eta}_{jk} | \mathbf{x}_{jk}, \mathbf{x}_k) f_{\theta}(\boldsymbol{\zeta}_{jk} | \mathbf{x}_{jk}, \mathbf{x}_k) d\boldsymbol{\eta}_{jk} d\boldsymbol{\zeta}_{jk}.
\end{aligned} \tag{23}$$

In the next step, one requires the assumption of conditional independence again, this time of the independence of  $N_k$  level-one response vectors nested within level-two unit  $k$ , conditionally on all the level-two latent variables and the background covariates:

$$\begin{aligned}
& f_{\theta}(\{\mathbf{y}_{jk}\}_{j=1}^{N_k} | \{\mathbf{x}_{jk}\}_{j=1}^{N_k}, \mathbf{x}_k) \\
&= \iint \left[ \int \prod_{j=1}^{N_k} f_{\theta}(\mathbf{y}_{jk} | \boldsymbol{\xi}_k, \boldsymbol{\eta}_k, \mathbf{x}_{jk}, \mathbf{x}_k) f_{\theta}(\boldsymbol{\xi}_k | \boldsymbol{\zeta}_k) d\boldsymbol{\xi}_k \right] f_{\theta}(\boldsymbol{\zeta}_k | \mathbf{x}_k) f_{\theta}(\boldsymbol{\eta}_k | \mathbf{x}_k) d\boldsymbol{\zeta}_k d\boldsymbol{\eta}_k,
\end{aligned} \tag{24}$$

where the parenthetical notation  $\{\mathbf{y}_{jk}\}_{j=1}^{N_k}$  and  $\{\mathbf{x}_{jk}\}_{j=1}^{N_k}$  simply collects together all the level-one outcome and covariate observations within level-two unit  $k$ .

Suffice it to say that Equation (24) contains the marginal likelihood function of all parameters of the model, if the observations are treated as fixed. Summing across the (assumed) independent level-two units, the overall marginal log-likelihood function is:

$$\log L(\boldsymbol{\theta} | \mathbf{Y}, \mathbf{X}) = \sum_{k=1}^K \log f_{\theta}(\{\mathbf{y}_{jk}\}_{j=1}^{N_k} | \{\mathbf{x}_{jk}\}_{j=1}^{N_k}, \mathbf{x}_k). \tag{25}$$

Maximization of the log-likelihood typically requires numerical integration. A number of computational algorithms may be used, including the venerable EM algorithm with fixed quadrature due to Bock and Aitkin (1981), or the more recent stochastic approximation based Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2010bc). A comparative review involving both EM and MH-RM algorithms are provided by Cai and Thissen (2014) and newer IRT software programs such as IRTPRO (Cai, Thissen, & du Toit, 2011) and flexMIRT® (Houts and Cai, 2015) implement a large subset of the full features of the generalized modeling framework. Alternatively, one may also adopt fully Bayesian estimation for maximal flexibility in model specification and parameter estimation (see Edwards, 2010 for a review in the context of IRT).

As mentioned earlier, the assumption of conditional independence is salient in the modeling framework described above. The degree to which subsequent model-based

inferences are valid hinges on whether conditional independence is tenable. In practice, it is critical to check whether the latent variable dimensionality is correctly specified. In discussions of IRT model fit evaluation in Section 5.2, statistics that are sensitive to dimensionality misspecification are discussed.

## 4 Illustrative Applications

The general model presented in Section 3 is not entirely identified without further constraints and specialization. For example, in typical applications, the distribution of latent variables and disturbance terms will contain some fixed means and covariance components in order to identify the location and scale of the unobserved variables in the model. Multidimensional IRT models also require sufficient constraints in order to avoid rotation and reflection indeterminacy. What will be presented next is a set of applications, partly as continuations of the motivating examples from earlier discussions, and partly to illustrate the broad applicability of the modeling framework reviewed here. They are not meant to be exhaustive.

### 4.1 Two-tier Item Factor Model for Large-Scale Educational Assessments

Full-information item factor analysis (Bock, Gibbons, & Muraki, 1988) has become an important tool in large-scale educational assessment research. A special model, item bifactor model (Gibbons & Hedeker, 1992; Gibbons et al., 2007; Cai, Yang, & Hansen, 2011), has also increasingly captured psychometricians' attention (Reise, 2012). In a bifactor model, there is one primary factor, representing a target construct being measured, and there are  $S$  group-specific factors that are conditionally independent given the general factor. An item may load on at most one group-specific factor. The bifactor pattern is an example of a *hierarchical* solution (Holzinger & Swineford, 1937; Schmid & Leiman, 1957), dating back to earlier days in the factor analysis tradition.

Cai's (2010a) two-tier item factor model is a more general version of the bifactor model in that the number of general factors can be more than one. In a two-tier model, two kinds of latent variables are considered, primary  $\boldsymbol{\eta}_G$  and group-specific  $\boldsymbol{\eta}_S$ . In most cases, the group-specific factors are mutually orthogonal, and also orthogonal to the primary dimensions as an identification restriction (Rijmen, 2009). In a two-tier model, an item may load on all

primary factors, subject to model identification (such as rotational indeterminacy), and on at most 1 group-specific factor.

Figure 2 represents a hypothetical two-tier model with ten items (the rectangles) that load on 2 correlated primary factors, as well as 4 group-specific factors that are mutually orthogonal. In terms of the factor pattern, the 10 by 6 factor pattern matrix corresponding to the model in Figure 2 has the following form:

$$\begin{pmatrix} \beta & & & & & & \beta^* & & & & \\ \beta & & & & & & \beta^* & & & & \\ \beta & & & & & & \beta^* & & & & \\ \beta & & & & & & & \beta^* & & & \\ \beta & \beta & & & & & \beta^* & & & & \\ \beta & \beta & & & & & & \beta^* & & & \\ & & \beta & & & & & \beta^* & & & \\ & & \beta & & & & & \beta^* & & & \\ & & \beta & & & & & & \beta^* & & \\ & & \beta & & & & & & & \beta^* & \\ & & \beta & & & & & & & & \beta^* \end{pmatrix},$$

where nonempty entries indicate free parameters, and  $\beta$  simply denotes a free slope parameter and does not imply equality constraints. The importance of the two-tier or item bifactor constraints is also a computational one. Maximum marginal likelihood estimation of the two-tier model requires at most one additional fold of integration than the number of primary factors. Substantial computational time savings can be accomplished if software programs take advantage of this fact.

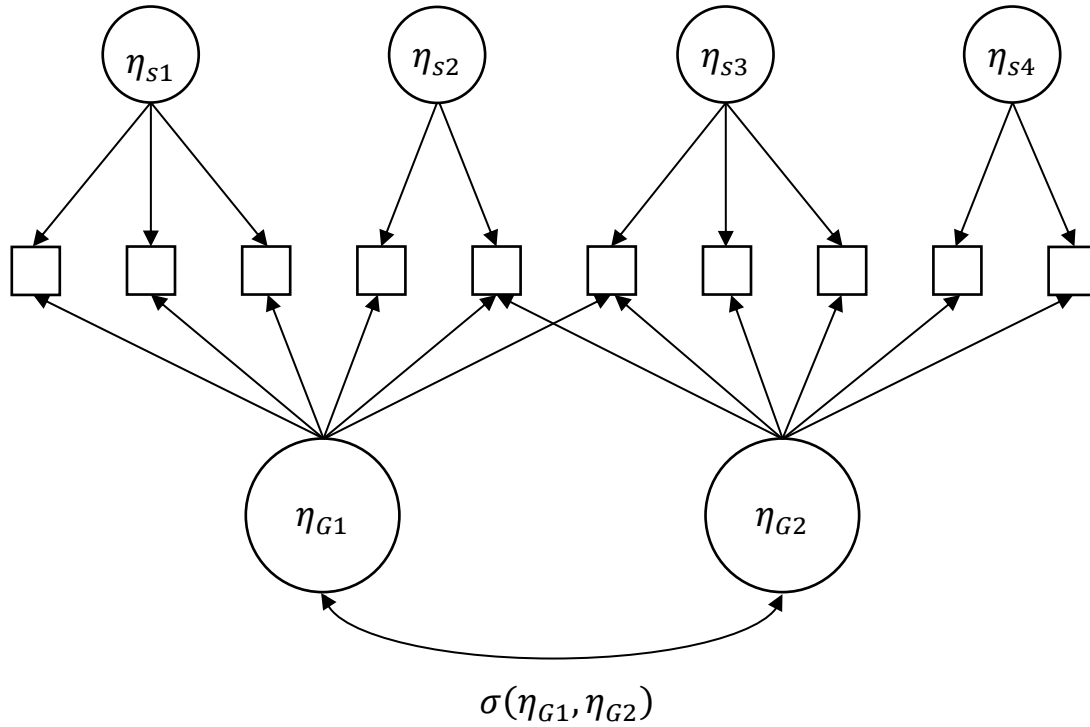


Figure 2. A two-tier model with 2 correlated primary factors and 4 group-specific factors

The two-tier model may also be useful for testlet-based educational assessments. Take the PISA reading literacy assessment as a case in point. The test design (see, e.g., Adams & Wu, 2002) calls for several potentially correlated “reading process” dimensions (subscales) that are thought to contribute to the domain specification of reading literacy. These subscales are interpretation, reflection/evaluation, and retrieval of information. In addition, the format of the assessment utilizes passage-based reading tests with several questions following each prompt (forming into testlets). This design introduces extra dependence among the items in the same testlet because they all depend on the specific content of the same passage, which should be accounted for in IRT modeling.

Figure 3 shows a potential model for such an assessment. The 3 primary factors correspond to the three reading process subscales. Items in the same testlet can belong to different subscales, and the total number of group-specific factors is equal to six, one for each task. The two-tier model is accounting for multidimensionality/dependence within and/or between the correlated subscales. Furthermore, though not shown in the figure, it is possible to regress the reading process factors on background covariates from the PISA

context questionnaire, e.g., student's academic interest and motivation factors. Harrell (2015) fitted such a model using NAEP data and the MH-RM algorithm.

Cai (2010a) fitted two-tier models to PISA reading and math assessment data simultaneously and not only found that the model offered improved fit, and can account for the substantial effect of residual item dependence within testlets, but were also able to offer improved estimates of target dimensions of interest. This is because the two-tier item factor model can utilize the estimated factor correlations (often substantial across domains or subdomains) to “borrow strength” from other the dimensions in the model in producing more precise marginal estimates of the latent variables. This is a familiar phenomenon in the context of shrinkage and subscore computations (Wainer et al., 2001; Haberman, 2008).

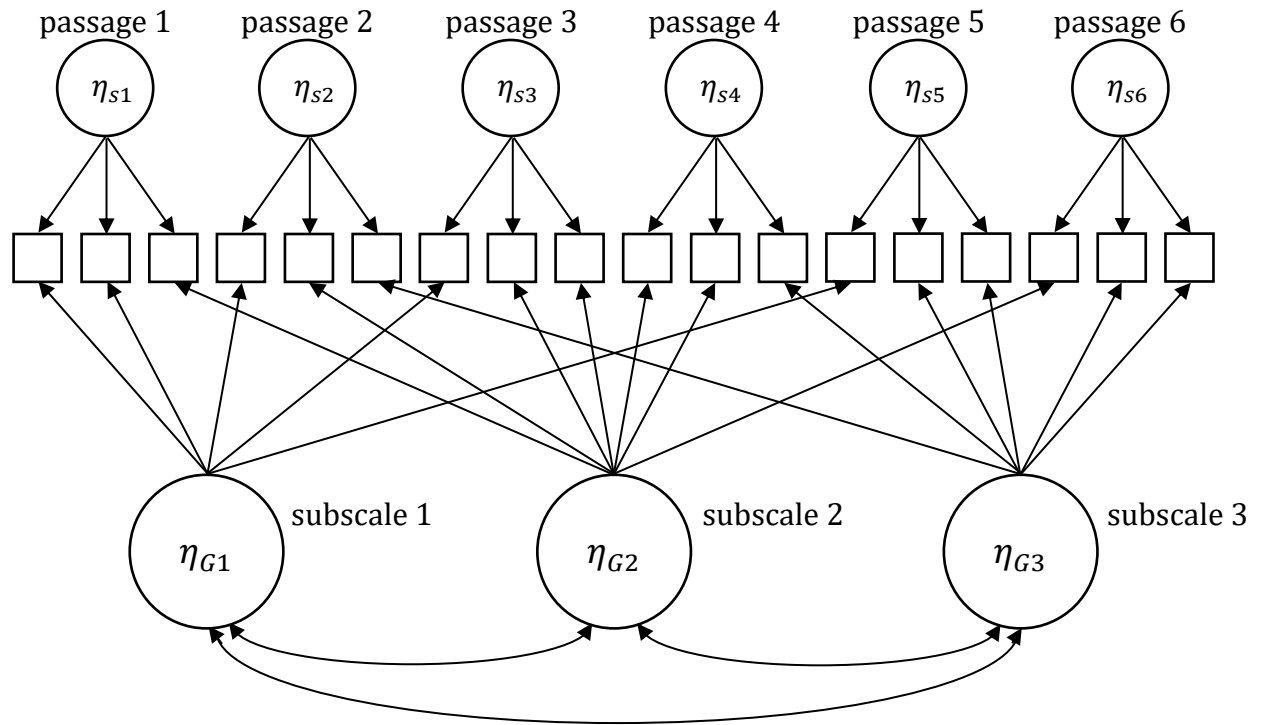


Figure 3. A two-tier model with potential application to testlet-based assessment

#### 4.2 Measurement and Efficacy of Intervention Studies

Presented here is an application of a multilevel two-tier item factor model to a large-scale randomized trial examining the instructional effects of serious video games with over 1500 students in 30 intervention classrooms and 29 comparison classrooms in 26 schools in 9

districts (Chung et al., 2015). The schools are the experimental sites and intact classrooms within schools are randomly assigned. These serious learning games aim to improve students' knowledge of pre-algebra topics such as rational numbers and fractions. Students' mathematics learning outcomes were measured by items similar to pre-algebra standardized assessment items (e.g., what might be included in state end-of-year tests), both before and after the experiment. Students in intervention classrooms played games on the topic of rational numbers and fractions, whereas those in the comparison classrooms played an alternative set of games on unrelated topics. This design is representative of a large class of evaluation designs used in federally funded education research. The hypothesis of the research team is that the students in intervention classrooms did better overall than their peers in the control classrooms, when pre-test differences are controlled for.

Without IRT, let  $Y_{ij}$  denote the observed summed score from student  $i$  in school  $j$  on the pretest and posttest assessments. The traditional hierarchical linear model (HLM; Raudenbush & Bryk, 2002) is widely used to estimate the impact of the intervention. The observed pretest summed score is entered into the model as a covariate. The pretest scores are school-mean centered so the random coefficient  $\beta_{0j}$  could represent the mean posttest score for school  $j$ . The parameter of interest,  $\beta_{1j}$  is the expected difference in posttest means between the two groups in school  $j$ , holding pretest constant. The HLM may be compactly represented as:

$$Y_{ij} = \beta_{0j} + \beta_{1j} * Trt_{ij} + \beta_{2j} * Pretest_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \quad (26)$$

$$\beta_{0j} = \gamma_{00} + u_{0j}, \quad u_{0j} \sim N(0, \tau_{00})$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \quad u_{1j} \sim N(0, \tau_{11})$$

$$\beta_{2j} = \gamma_{20} + u_{2j}, \quad u_{2j} \sim N(0, \tau_{22})$$

The fixed regression coefficients ( $\beta$ ) and the variance components ( $\tau$ ) are estimated directly using maximum likelihood. The overall treatment impact estimate is focused on the regression parameter  $\gamma_{10}$ , which is estimated to be 1.12 ( $SE = .33$ ). The posttest standard deviation is 4.99, and one can convert the unstandardized regression coefficient into an estimated standardized effect size (Cohen's  $d = 1.12/4.99 = .22$ ). While it represents a statistically significant effect, the standardized effect size would be considered small to

moderate in size. A concern the research team had was that the assessments were not long and may contain substantial degree of measurement error. Because the assessment scores appear on both sides of the regression equation, the question is whether the results can be improved by applying more modern methods of data analysis such as IRT.

As mentioned earlier, classical summed scores (or standard IRT scaled scores) have a number of limitations when used in conjunction with cluster randomized multisite experimental studies. Cai (2010a) noted that in calibrating IRT model with longitudinal item response data, even if the measurement instrument may be unidimensional at each time point, the longitudinal item response data are inherently multidimensional. For designs with pretest and posttest, at least two occasion-specific primary latent factors are needed for each of the treatment or control groups to model the initial status and potential gains. In addition, the responses to the same item in pretest (time 1) and posttest (time 2) from the same individual may be residually dependent, even after controlling for the influence of the primary dimensions. Thus, item-specific residual dependence factors (doublets) are introduced to handle the potential residual dependence, and there are as many of them as the number of repeated items.

Multisite randomized experiments also generally consist of two or more groups—control and treatment in this case—within each site. Thus, it is necessary to specify four within-site primary dimensions to represent the pretest and post status of treatment and control groups. This is akin to conditioning the latent outcome variable on the treatment assignment indicator. Finally, from the study design, it is clear that multisite data have a nested structure, thus requiring both level-one and level-two (site) latent variables to model the site-level variation in initial status and change due to treatment receipt.

For a generic item  $i$  that appeared in both pretest and posttest, the linear predictor portions of the item response models could be written as the following:

$$\begin{aligned}
 \text{Pretest Control: } & \beta_i[\eta_{1k} + \eta_{1jk}] + \beta_i^* \eta_{ijk}^* \\
 \text{Posttest Control: } & \beta_i[(\eta_{1k} + \eta_{1jk}) + (\eta_{2k} + \eta_{2jk})] + \beta_i^* \eta_{ijk}^* \\
 \text{Pretest Treatment: } & \beta_i[\eta_{1k} + \eta_{3jk}] + \beta_i^* \eta_{ijk}^* \\
 \text{Posttest Treatment: } & \beta_i[(\eta_{1k} + \eta_{3jk}) + (\eta_{2k} + \eta_{4jk})] + \beta_i^* \eta_{ijk}^*
 \end{aligned} \tag{27}$$

Each item has an overall discrimination parameter  $\beta_i$  and the various latent variables contribute to the item response in a systematic manner. Specifically, two between-school latent variables,  $\eta_{1k}$  and  $\eta_{2k}$  represent latent initial status and latent gain between pretest and posttest for school  $k$ , respectively. In addition, among the four within-school latent variables, the first two latent variables represent initial status ( $\eta_{1jk}$ ) and latent gain ( $\eta_{2jk}$ ) for student  $j$  in the control condition within school  $k$ , and the rest represent initial status ( $\eta_{3jk}$ ) and latent gain ( $\eta_{4jk}$ ) for student  $j$  in the treatment condition within school  $k$ . As such, the additional latent variables at posttest represent potential gains over the pretest level. Furthermore, the variation is decomposed at both pretest and posttest into between-school and within-school components. By allowing mean differences between  $\eta_2$  and  $\eta_4$  be estimated, that is, the difference between latent changes between the treatment and control conditions within schools, potential effects of treatment on the *gain* in learning may be explicitly represented. The term  $\eta_{ijk}^*$  is an item-specific random effect that accounts for the residual dependence over repeated administrations of the same item in pretest and posttest.

This model is motivated by growth modeling developments as represented in Bock and Bargmann (1966), Embretson (1991), and McArdle (2009), among others. Upon estimating the item parameters, proper IRT scaled scores can then be computed for each of the latent variables as posterior means (Expected *A Posteriori* (EAP) scores, Thissen & Wainer, 2001) or as multiple imputations (plausible values; see von Davier, Gonzalez, & Mislevy, 2009).

The same HLM as in Equation (26) may be used to estimate the impact of the intervention. This time, the standard effect size in terms of Cohen's  $d$  is  $.24/.42 = .57$ . The substantial increase in estimated effectiveness stems from the fact the latent variables serve to isolate that part of the posttest individual differences in learning outcomes that is sensitive to change, after controlling for prior knowledge. In contrast, the standard approach of using posttest scores (observed or scaled latent variable estimates) as an outcome conflates two sources of variation, i.e., prior knowledge and potentially malleable difference. As a result, the latent gain score dimensions from the MTT model are more



sensitive measures for targeted interventions than the observed outcome scores (for a similar phenomenon, see Gibbons et al., 2008, p. 365).

Multilevel IRT models may also be used in observational studies. For instance, Yang & Cai (2014) examined the application of multilevel IRT models to the study of contextual effects. The estimation of these multilevel IRT models are non-trivial and stochastic optimization algorithms are often necessary.

### 4.3 Improving Diagnostic Measurement

This example builds on prior work by Lee, Park, & Taylan (2011), who analyzed data from booklets 4 and 5 from the 2007 Trends in Mathematics and Science Study (TIMSS) fourth grade mathematics test. They had 25 test items reviewed and rated according to the specific testing objectives described in the assessment framework. Of these items, 15 unique testing objectives were identified. Accordingly, they set up a  $\mathbf{Q}$ -matrix with 25 rows and 15 columns. The attributes represent the skills required in each of the testing objectives.

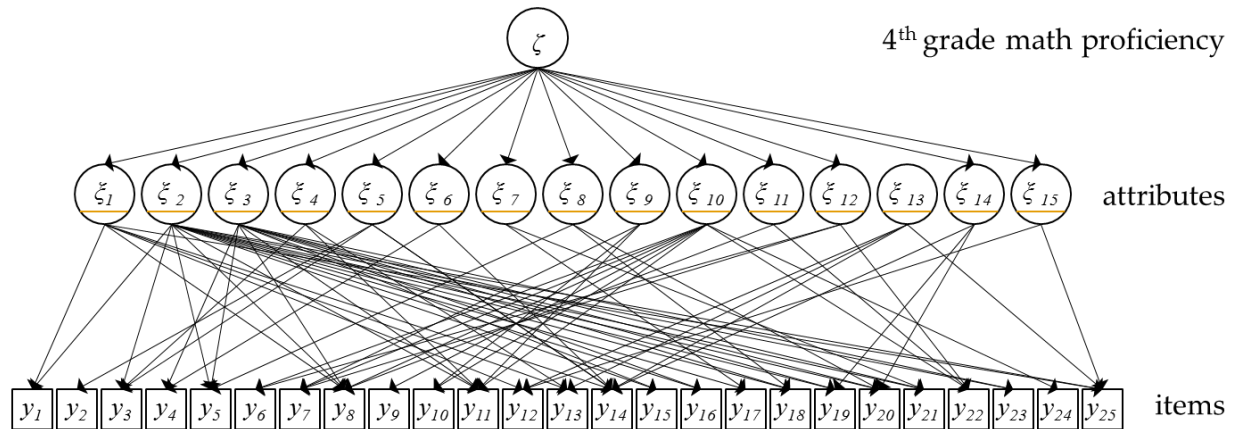


Figure 4. Path diagram of original  $\mathbf{Q}$ -matrix and higher-order model structure.

Lee et al. (2011) specified a conjunctive DINA model for the items in their analysis, i.e., skills in all testing objectives measured by an item will have to be mastered for a correct solution. A higher-order DINA model (de la Torre & Douglas, 2004) using the reported  $\mathbf{Q}$ -matrix was specified and fitted to a sample of 564 students from the US. Figure 4 presents a path diagram of this higher-order DCM. The latent attribute variables are shown as circles with a line across to indicate that they are discrete. The attribute variables are regressed further on a single higher-order factor that may be thought of as math proficiency. The item

to attribute relationship is complex, with most items associated with a number of attributes. This is fairly typical of DCMs because the attributes represent narrow skills, in contrast to item factor analysis models that contain factors representing broad traits or abilities.

The items considered in this study have all since been made public (Foy & Olson, 2009) and were thus available for review. Upon fitting the model, it was clear that the model failed to completely account for the dependence between items 18 and 19. Reviewing the content of the items showed that the two items are administered within a testlet (M031242A/B/C), and it was evident upon further inspection that a correct response to item 18 (M031242A) could greatly help answering item 19. This observation may be the reason why the initial DCM could not explain the dependence between these two items. Adding another latent factor to explain the residual dependence of the item doublet results in the addition of a single parameter. This is akin to adding a random intercept as in Wainer, Bradlow and Wang's (2007) testlet response theory model. A likelihood ratio test between the initial model and the modified model shows that the addition of the extra parameter led to significant model fit improvement  $\Delta\chi^2(1) = 53, p < .001$ .

## 5 Estimation and Model Fit

### 5.1 IRT Parameter Estimation

As Cai and Thissen (2014) mentioned in their recent review on IRT parameter estimation methods, a central idea that runs through parameter estimation in IRT is the missing data formulation, wherein the latent variables are regarded as missing observations that should be filled in. Once they are properly filled in, the remainder of parameter estimation becomes straightforward because one can take advantage of the conditional independence structure inherent in the model.

This can be illustrated with an example. For simplicity, consider a unidimensional IRT model. Given the foregoing discussions, the marginal probability associated with response pattern  $\mathbf{y}_j = (y_{1j}, \dots, y_{nj})$  is

$$f_{\theta}(\mathbf{y}_j) = \int \prod_{i=1}^n \prod_{c=0}^{C_i-1} T_i(c|\eta)^{\chi_c(y_{ij})} f(\eta) d\eta = \int f_{\theta}(\mathbf{y}_j, \eta) d\eta. \quad (28)$$

If the item responses  $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$  are regarded as fixed once they are observed, the marginal likelihood function for all the item parameters in  $\boldsymbol{\theta}$  can be expressed as:

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{j=1}^N \int \prod_{i=1}^n \prod_{c=0}^{C_i-1} T_i(c|\eta)^{\chi_c(y_{ij})} f(\eta) d\eta. \quad (29)$$

Because the marginal likelihood  $L(\boldsymbol{\theta}|\mathbf{Y})$  does not depend on the unobserved  $\eta$  values, it may be referred to as the *observed* data likelihood. The observed data likelihood can be approximated via numerical quadrature and directly optimized using Newton-type algorithms (Moustaki & Knott, 2014) or the EM algorithm (Bock & Aitkin, 1981).

Implicit in the “observed data” terminology is a realization that the latent variables are the missing data. Thus, the so-called “complete data” likelihood function is

$$L(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\eta}) = \left[ \prod_{j=1}^N f(\eta_j) \right] \left[ \prod_{j=1}^N \prod_{i=1}^n \prod_{c=0}^{C_i-1} T_i(c|\eta)^{\chi_c(y_{ij})} \right], \quad (30)$$

where  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_N)$  collects together all the latent variable scores. Equation (30) shows that the complete data likelihood takes a factored form because of the assumption of item conditional independence and the assumed independence of individuals. It is also easy to see that the complete data likelihood is proportional to the posterior distribution of  $\boldsymbol{\eta}$ ,  $p_{\boldsymbol{\theta}}(\boldsymbol{\eta}|\mathbf{Y})$ . With the help of this posterior distribution, one may verify that given item parameter values, the following equality holds (it is known as Fisher’s Identity; Fisher, 1925) assuming mild regularity conditions:

$$\frac{\partial \log f_{\boldsymbol{\theta}}(\mathbf{y}_j)}{\partial \boldsymbol{\theta}} = \int \frac{\partial \log f_{\boldsymbol{\theta}}(\mathbf{y}_j, \eta)}{\partial \boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\eta|\mathbf{y}_j) d\eta. \quad (31)$$

In other words, the gradient of the observed data likelihood is equal to the expectation of the complete data likelihood over the posterior distribution of the latent variables given the observed variables. This insight suggests that instead of maximizing the observed data likelihood, one could iteratively maximize the conditional expected complete data likelihood, eventually ending up at the same solution.

Bock and Aitkin's (1981) solution can be considered a direct application of Fisher's Identity. They noted that the marginal probability can be approximated by replacing the integral with a summation over a set of quadrature points spread over  $\eta$ :

$$f_{\theta}(\mathbf{y}_j) \approx \bar{P}_j = \sum_{s=1}^S \prod_{i=1}^n \prod_{c=0}^{C_i-1} T_i(c|A_s)^{\chi_c(y_{ij})} W_s, \quad (32)$$

where  $A_s$  is a quadrature point, and  $W_s$  is the corresponding weight, e.g., from Gauss-Hermite integration formulas. Bock and Aitkin (1981) also saw that the height of the posterior distribution at quadrature point  $A_s$  could be approximated:

$$p_{\theta}(A_s|\mathbf{y}_j) \approx \frac{\prod_{i=1}^n \prod_{c=0}^{C_i-1} T_i(c|A_s)^{\chi_c(y_{ij})} W_s}{\bar{P}_j}. \quad (33)$$

Ignoring constants involving the prior distribution  $f(\eta_j)$  from Equation (30), the complete data log-likelihood for the item parameters can be written as

$$\log L(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\eta}) \propto \sum_{j=1}^N \sum_{i=1}^n \sum_{c=0}^{C_i-1} \chi_c(y_{ij}) \log T_i(c|\eta). \quad (34)$$

Thus, the conditional expected complete data log-likelihood given provisional item parameter values  $\boldsymbol{\theta}^*$  can be approximated as follows

$$Q(\boldsymbol{\theta}|\mathbf{Y}; \boldsymbol{\theta}^*) \approx \sum_{j=1}^N \sum_{s=1}^S \sum_{i=1}^n \sum_{c=0}^{C_i-1} \chi_c(y_{ij}) \log T_i(c|A_s) p_{\theta^*}(A_s|\mathbf{y}_j). \quad (35)$$

The third important insight from Bock and Aitkin (1981) is that by interchanging the order of summation, they realized that the posterior probabilities can be accumulated over individuals first:

$$Q(\boldsymbol{\theta}|\mathbf{Y}; \boldsymbol{\theta}^*) = \sum_{i=1}^n \sum_{s=1}^S \sum_{c=0}^{C_i-1} r_{isc} \log T_i(c|A_s), \quad (36)$$

where  $r_{isc} = \sum_{j=1}^N \chi_c(y_{ij}) p_{\theta^*}(A_s|\mathbf{y}_j)$  is the conditional expected probabilities of individuals in category  $c$  for item  $i$  at quadrature point  $A_s$ . As one can see, the conditional expected complete data model in Equation (36) is far more tractable than the observed data model in Equation (29) because the complete data model reduces to a series of independent logit analyses with  $A_s$  as the predictor values and  $r_{isc}$  as the outcome weights. In the E-step, the

conditional expected probabilities are computed. In the M-step, the conditional expected complete data log-likelihood is maximized. The two steps are repeated until convergence.

While the Bock-Aitkin EM algorithm elegantly handles unidimensional and sufficiently low-dimensional IRT parameter estimation, it does not generalize to the case of more complex IRT models such as the kind discussed in Section 3. This is because even with a moderate number of latent dimensions, the exponentially increasing size of the quadrature grid presents an insurmountable computational burden so far. Adaptive quadrature (see, e.g., Schilling & Bock, 2005) has been employed, to limited but not complete success. Wirth and Edwards (2007) referred to the problem as the “challenge of dimensionality”.

Cai’s (2010bc) noted that a solution already resides in Fisher’s Identity. If one could draw multiple random imputations of  $\eta$  from its posterior predictive distribution  $p_{\theta^*}(\eta|\mathbf{y}_j)$ , with provisional item parameter estimates  $\theta^*$ , the right hand side of Equation (31) can be approximated by Monte Carlo, which is the same as the left hand side over repeated sampling. In other words, the Monte Carlo average of complete data log-likelihood gradients gives the same likelihood ascent direction as the observed data log-likelihood gradient vector! With the widely spread availability of Markov chain Monte Carlo (MCMC) methods (e.g., the Metropolis-Hastings methods; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970), the imputations can be produced rather easily because the complete data likelihood is proportional to the posterior distribution of  $\eta$ .

One problem still remains, and that is the Monte Carlo approximation contains error, and unless the Monte Carlo size increases, the random sampling error obscures the true direction of likelihood ascent. This is a known issue in the context of Monte Carlo EM (Booth & Hobert, 1999). Cai (2010b) noted that instead of trying to contain the Monte Carlo noise as a nuisance, it may in fact be employed productively. By drawing an analogy to the classical Robbins and Monro’s (1951) Stochastic Approximation (SA) algorithm, the Monte Carlo noise can be regarded as the stochastic excitations that drive an underlying dynamical system with the maximum marginal likelihood solution a stability point. The noise can be filtered out step-by-step with the use of a sequence of slowly decaying gain constants. This results in the Metropolis-Hastings Robbins-Monro algorithm that borrows elements of

MCMC and combines them with SA. The MH-RM algorithm can be used to estimate all of the models described here and because it eschews quadrature, it is computationally manageable even for high-dimensional IRT models.

## 5.2 IRT Model Fit Evaluation

Upon fitting the model, the issue of model fit evaluation becomes immediately relevant. Given the fact the models discussed here are highly parameterized and contain a large number of latent variables, also making rather strict assumptions related to distributional form and conditional independence, it is critical that overall model fit tests and diagnostic indices are developed. Unfortunately, this is an area wherein IRT modeling has historically lagged significantly behind other latent variable modeling frameworks (e.g., structural equation model). In recent years, however, two strands of model fit evaluation research emerged within IRT, providing promising tools that are being increasingly utilized in the data analytic practice.

The first strand involves limited-information goodness of fit testing (see Maydeu-Olivares & Joe, 2005; Cai, Maydeu-Olivares, Coffman & Thissen, 2006; Maydeu-Olivares, 2013 and the references therein). The IRT model is built on an underlying multiway contingency table made up of the full item  $\times$  item  $\times$  item ... cross classifications. For dichotomously scored items, the size of this contingency table grows exponentially as the number of items increases ( $2^n$ ). With a modest number of items, e.g., 20, the total number of response pattern probabilities in the underlying general multinomial easily surpasses the sample size. Therefore, classical likelihood ratio or Pearson chi-square statistics no longer have their purported asymptotic chi-square null distributions due to the issue of sparseness (Bartholomew & Tzamourani, 1999), though conditional inference based on these statistics remain useful (McCullagh, 1986). A more recent solution advocated by IRT researchers in recent years is to examine the lower-order marginal tables of the underlying multinomial because the cells in these lower-order (e.g., first- and second-order) margins are much better filled. Wald-type chi-squares can then be constructed (see e.g., Cai & Hansen, 2013). Despite an apparent loss of information (because the higher-order moments are ignored), the limited-information fit statistics do have much better calibration than the classical (full-

information) counterparts and can also be more powerful against important alternatives that impact the lower-order marginal probabilities such as dimensionality misspecifications. One advantage of the limited-information model fit tests is that the test statistics can be easily decomposed into univariate and bivariate residuals (see e.g., Liu & Maydeu-Olivares, 2013, 2014) so that the sources of model misfit can be examined.

The second strand involves posterior predictive model checking (PPMC; Glas & Meijer, 2003; Hoijtink & Molenaar, 1997; Janssen, Tuerlinckx, Meulders, & de Boeck, 2000; Levy, Mislevy, & Sinharay, 2009; Levy & Svetina, 2011; Sinharay, Johnson, & Stern, 2006). In PPMC, instead of relying on classical asymptotic arguments and chi-square indices, the researchers took the alternative stance that model fit evaluation can be determined on the basis of comparisons of realized vs. observed discrepancies from repeated simulations of plausible (future) data sets under the fitted model. This inherently Bayesian concept (Rubin, 1981, 1984) has penetrated IRT such that it has already been applied to the assessment of in assessing item-level model fit, person fit, and dimensionality. An interesting connection between the PPMC literature and the more classical model evaluation literature is in the choice of discrepancy measures. In PPMC, choices of discrepancy measures lead to differentially diagnostic indices sensitive against different forms of misspecification. Some classically derived indices such as Orlando and Thissen's (2000) summed score based item fit indices have been found successful as discrepancy measures.

## 6 Discussion

After decades of development, IRT has become a staple in applied statistics in the social and behavioral sciences. This is further exemplified by the recent emergence of dedicated IRT procedures in large statistical software packages such as Stata and SAS. A plethora of R packages also exist (e.g., mirt; Chalmers, 2012) that often offer enhanced features found in more recent research literature. The application of IRT has also expanded quite considerably from its original base in educational measurement. The diversification adds value to IRT but also makes this review difficult. For instance, a number of IRT models have not been covered.

*Response time models.* With the proliferation of computer-based data collection, individual response time data are increasingly collected along with the responses themselves. Though the item-specific response time outcomes are no longer discrete data, response time data can be integrated into the IRT modeling framework and jointly modeled with item responses. Response time models of log-normal type (see e.g., van der Linden, 2006) have since been developed to account for the time-intensity of items and persons' speed. Response time models have also been used in IRT-based computerized adaptive testing (van der Linden, 2008).

*Crossed random effects.* The IRT models discussed here are nested random effects models. There is an altogether different class of IRT models that regards the items as crossed with individuals (De Boeck, 2008). In these models, item parameters are no longer fixed, but are random variables themselves. Such a conceptual shift leads to interesting generalizability potential of IRT models (Briggs & Wilson, 2007). One significant application of crossed random effect IRT models is in the area of automatic item generation (Embretson, 1999; Glas & van der Linden, 2003; Sinharay, Johnson, & Williamson, 2003; Janssen, Tuerlinckx, Meulders, & De Boeck, 2000).

*Special Response Processes.* The IRT models discussed in this review contain relatively complex latent structural models but simple logit-based link functions. Models for special response processes exist. First, semi-parametric or non-parametric item response functions have recently been considered (Falk & Cai, in press; Liang, 2007; Miyazaki & Hoshino, 2009) as a compromise between older approaches such as kernel smoothing (Ramsay, 1991) and even older and more restrictive parametric ones reviewed here. Related, there are also IRT models with semi-parametric latent distributions (e.g., Monroe & Cai, 2014; Woods & Thissen, 2006). Second, there is another class of models that generalize the assumption of monotonic item response functions (dominance response process) - the so-called ideal point (unfolding) models (e.g., Andrich, 1996). In unfolding models item response functions have peaks, reflecting ideal points of response along the latent dimension. There has also been some interesting developments in multidimensional unfolding models (e.g., Maydeu-Olivares, Hernández, McDonald, 2006). Third, the IRT models discussed thus far are not



appropriate for forced-choice questionnaires. In forced-choice items, respondents have to make choices between two or more items presented at the same time. This introduces opportunities for improved measurement but also induces complex dependencies in the data. Brown (in press) offers a general framework for analyzing data arising from forced-choice items. Finally, in assessment contexts other than educational testing, individuals can and often do exhibit response styles, e.g., middle response, socially desirable response, extreme response, etc. Recently formal models have been developed in IRT to handle the extra complexities induced by individual response styles (Böckenholt, 2012, 2014; Falk & Cai, in press; Khorramdel & von Davier, 2014; Plieninger & Meiser, 2014; Thissen-Roe & Thissen, 2013), with emerging utility for non-traditional applications of IRT.

Because of IRT's long history, the purpose of this review is a rather modest one. What has been offered is a general statistical framework based on recent technical developments in multidimensional and multilevel IRT modeling, along with a conceptual map that tries to unite several sub-strands of research within IRT. It is fair to conclude that the statistical field of IRT modeling will continue to enjoy active development in the coming decades as the cost of data collection drops further in the social and behavioral sciences, and that technology enhances the role of evidence-based assessment and evaluation in decision-making.

## 7 References

- Adams, R. J., & Wu, M. L. (Eds.) (2002). *PISA 2000 technical report*. Paris: OECD Publications.
- Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology*, 49, 347-365.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. New York: Dekker.

- Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Bock, R. D. (1997). A Brief History of Item Theory Response. *Educational Measurement: Issues and Practice*, 16, 21-33.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Bock, R. D., & Bargmann, R.E. (1966). Analysis of covariance structures. *Psychometrika*, 31, 507-533.
- Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, 12, 261-280.
- Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*. 17, 665-678.
- Böckenholt, U. (2014). Modeling motivated misreports to sensitive survey questions. *Psychometrika*. 79, 515-537.
- Booth, J. G., & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society – Series B*, 61, 265–285.
- Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, 44(2), 131-155.
- Brown, A. (in press). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*.
- Cai, L. (2010a). A two-tier full-information item factor analysis model with applications. *Psychometrika*, 75, 581-612.
- Cai, L. (2010b). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, 75, 33-57.

- Cai, L. (2010c). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35, 307-335.
- Cai, L. (2015). Lord-Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika*, 80, 535-559.
- Cai, L., Choi, K., & Kuhfeld, M. (in press). On the role of multilevel item response models in multi-site evaluation studies for serious games. In H. F. O'Neil, E. L. Baker, R. Perez. (Eds.), *Issues regarding the use of games and simulations for teaching and assessment*. New York, NY: Taylor & Francis.
- Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, 66, 245-276.
- Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited information goodness-of-fit testing of item response theory models for sparse  $2^p$  tables. *British Journal of Mathematical and Statistical Psychology*, 59, 173-194.
- Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling* [Computer software]. Lincolnwood, IL: Scientific Software International.
- Cai, L., & Thissen, D. (2014). Modern approaches to parameter estimation in item response theory. In S. P. Reise & D. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. New York, NY: Taylor & Francis.
- Cai, L., Yang, J., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, 16, 221-248.
- Chalmers, R., P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48, 1-29.

- Chung, G. K. W. K., Choi, K., Baker, E. L., & Cai, L. (2014). *The effects of math video games on learning: A randomized evaluation study with innovative impact estimation techniques* (CRESST Report 841). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Choi, H.-J., Rupp, A. A., & Pan, M. (2013). Standardized diagnostic assessment design and analysis: Key ideas from modern measurement theory. In R. Maclean (Ed.), *Self-directed Learning Oriented Assessment in the Asia-Pacific*. New York: Springer.
- Curran, P., Hussong, A., Cai, L., Huang, W., Chassin, L., Sher, K., & Zuchek, R. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Developmental Psychology*, 44(2), 365-380.
- De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73, 533-559.
- de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.
- Edwards, M.C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, 75, 474-497.
- Embretson, S. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56, 495-515.
- Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, 64, 407-433.
- Falk, C. F., & Cai, L. (in press). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika*.
- Fisher, R.A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700-725.
- Foy, P. & Olson, J.F. (2009). *TIMSS 2007 user guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57(3), 423-436.
- Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., et al. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, 31, 4-19.
- Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, 59, 361-368.
- Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, 27, 217-233.
- Glas, C. A. W. & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, 27, 247-261.
- Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33, 204-229.
- Harrell, L. A. (2015). *Analysis strategies for planned missing data in health sciences and education research*. Unpublished doctoral dissertation, Department of Biostatistics, University of California, Los Angeles.
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.
- Hastings, W.K. (1970). Monte Carlo simulation methods using Markov chains and their applications. *Biometrika*, 57, 97-109.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

- Henson, R., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74, 191-210.
- Holzinger, K. J. & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2, 41-54.
- Houts, C.R., & Cai, L. (2013). *flexMIRT user's manual version 2.0: flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group.
- Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika*, 62, 171-180.
- Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*, 25, 285-306.
- Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258-272.
- Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the big five: a multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, 49(2), 161-177.
- Lazarsfeld, P. F. (1950) "The Logical and Mathematical Foundations of Latent Structure Analysis", Chapter 10 in Stouffer (1950).
- Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing*, 11, 144-177.
- Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, 33, 519-537.

- Levy, R., & Svetina, D. (2011). A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, 64, 208-232.
- Liang, L. (2007). *A semi-parametric approach to estimating item response functions*. Unpublished doctoral dissertation, Department of Psychology, The Ohio State University.
- Lindley, D. V., & Smith, A. F. M. (1972). *Journal of the Royal Statistical Society - Series B (Methodological)*, 34, 1-41.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley.
- Little, R. J. A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *The American Statistician*, 37, 218-220.
- Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, 73(2), 254-274.
- Liu, Y., & Maydeu-Olivares, A. (2014). Identifying the source of misfit in item response theory models. *Multivariate Behavioral Research*, 49, 354-371.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, 8, 453-461.
- Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, 11, 71-101.
- Maydeu-Olivares, A. & Joe, H. (2005). Limited and full information estimation and goodness-of-fit testing in 2n contingency tables: A unified framework. *Journal of the American Statistical Association*, 100, 1009-1020.
- Maydeu-Olivares, A., Hernández, A. & McDonald, R.P. (2006). A multidimensional ideal point IRT model for binary data. *Multivariate Behavioral Research*, 44, 445-472.
- McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *Journal of the American Statistical Association*, 81, 104-107.

- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state space calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087-1092.
- McArdle, J. J. (2009). Latent variable modeling of difference and changes with longitudinal data. *Annual Review of Psychology*, 60, 577-605.
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177-196.
- Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling Procedures in NAEP. *Journal of Educational Statistics*, 17, 131-154.
- Miyazaki, K., & Hoshino, T. (2009). A bayesian semiparametric item response model with dirichlet process priors. *Psychometrika*, 74(3), 375-393.
- Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-curve item response theory model by the Metropolis-Hastings Robbins-Monro algorithm. *Educational and Psychological Measurement*, 74, 343-369.
- Moustaki, I. & Knott, M. (2014). Latent variable models that account for atypical responses *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(2), 343-360.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50-64.
- Plieninger, H. & Meiser, T. (2014). Validity of multi-process IRT models for separating content and response styles. *Educational and Psychological Measurement*, 74(5), 875-899
- Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56(4), 611-630.



- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 4, 321–334.
- Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. 2nd edition. Newbury Park , CA : Sage.
- Raudenbush, S.W. & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5(2), 199-213.
- Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer.
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667-696.
- Reise, S. P. & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, 5, 27-48.
- Rijmen, F. (2009). *Efficient full-information maximum likelihood estimation for multidimensional IRT models*. (Technical Report No. RR-09-03). Princeton, NJ: Educational Testing Service.
- Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22, 400–407.
- Rubin, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151-1172.
- Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. J. Wiley & Sons, New York.
- Rupp, A. A., Templin, J. & Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press.
- Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22, 53-61.

- Schilling, S., & Bock, R.D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, 70, 533–555.
- Sinharay, S., Johnson, M. S., & Stern, H. S. (2006) Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30(4), 298-321.
- Sinharay, S., Johnson, M. S., & Williamson, D. M. (2003). *An application of a Bayesian hierarchical model for item family calibration* (ETS RR-03-04). Princeton, NJ: Educational Testing Service.
- Spybrook, J., & Raudenbush, S.W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, 31(3), 298-318.
- Stouffer, S. A., et al. (1950) *Measurement and Prediction*, Volume IV of *The American Soldier: Studies in Social Psychology in World War II*. Princeton University Press. Reprinted 1973 by Peter Smith, Gloucester MA.
- Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345–354.
- Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287-305.
- Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Development and applications* (pp. 43–75). New York, NY: Taylor & Francis.
- Thissen, D. & Wainer, H. (Eds) (2001) *Test Scoring*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Thissen, D. & Steinberg, L. (2009). Item response theory. In R. Millsap & A. Maydeu-Olivares (Eds.), *Handbook of quantitative methods in psychology*. London: Sage Publications.

- Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, 38, 522-547.
- Thomas, N. & Gan, N. (1997). Generating multiple imputations for matrix sampling data analyzed with item response models. *Journal of Educational and Behavioral Statistics*, 22(4), 425-446.
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16, 433-451.
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31, 181-204.
- van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, 33, 5-20.
- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS No. RR-05-16). Princeton, NJ: ETS.
- von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI Monograph Series*, 2, 9–36.
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.
- Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., & Nelson, L. (2001). Augmented scores—“Borrowing strength” to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Hillsdale, NJ: Lawrence Erlbaum.
- Wirth, R.J., & Edwards, M.C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12, 58-79.
- Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, 71, 281-301.

Yang, J. S., & Cai, L. (2014). Estimation of contextual effects through nonlinear multilevel latent variable modeling with a Metropolis-Hastings Robbins-Monro algorithm. *Journal of Educational and Behavioral Statistics*, 39, 550-582.